Chapter 1 - The Power of Statistical Tests

In the social and behavioral sciences, statistics serve two general purposes. First, they can be used to describe what happened in a particular study (descriptive statistics). Second, they can be used to help draw conclusions about what those results mean in some broader context (inferential statistics). The main question in inferential statistics is whether a result, finding, or observation from a study reflects some meaningful phenomenon in the population from which that study was drawn. For example, if 100 college sophomores are surveyed and it is determined that a majority of them prefer pizza to hot dogs, does this mean that people in general (or college students in general) also prefer pizza? If a medical treatment yields improvements in 6 out of 10 patients, does this mean that it is an effective treatment that should be approved for general use? The goal of inferential statistics is to determine what sorts of inferences and generalizations can be made on the basis of data of this sort, and to assess the strength of evidence and the degree of confidence one can have in these inferences.

The process of drawing inferences about populations from samples is a risky one, and a great deal has been written about the causes and cures for errors in statistical inference. Statistical power analysis (Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990) falls under this general heading. Studies with too little statistical power can lead to erroneous conclusions about the meaning of the results of a particular study. In the example cited above, the fact that a medical treatment worked for 6 out of 10 patients is probably insufficient evidence that it is truly safe and effective, and if you have nothing more than this study to rely on, you might conclude that the treatment has not been proven effective. Does this mean that you should abandon the treatment, or that it is unlikely to work in a broader population? The conclusion that the treatment has not been shown to be effective may say as much about the low level of statistical power in your study as about the value of the treatment.

In this chapter, we will describe the rationale for and applications of statistical power analysis. In most of our examples, we describe or apply power analysis in studies that assess the effect of some treatment or intervention (e.g., psychotherapy, reading instruction, performance incentives) by comparing outcomes for those who have received the treatment to outcomes of those who have not (non-treatment or control group). However, power analysis is applicable to a very wide range of statistical tests, and the same simple and general model can be applied to an virtually all of the statistical analyses you are likely to encounter in the social and behavioral sciences.

The Structure of Statistical Tests

To understand statistical power, you must first understand the ideas that underlie statistical hypothesis testing. Suppose 100 children are randomly divided in two groups. Fifty children receive a new method of reading instruction, and their performance on reading tests is on average six points higher (on a 100-point test) than the other 50 children who received standard methods of instruction. Does this mean that the new method is truly better? A six-point difference *might* mean that the new method is really better, but it is also possible that there is no real difference between the two methods, and that this observed difference is the result of the sort of random fluctuation you might expect when you use the results from a single sample to draw inferences about the effects of these two methods of instruction in the population.

One of the most basic ideas in statistical analysis is that results obtained in a sample do not necessarily reflect the state of affairs in the population from which that sample was drawn. For example, the fact that scores averaged six points higher in this particular group of children does not necessarily mean that scores will be six points higher in the population, or that the same six-point difference would be found in another study examining a new group of students. Because samples do not (in general) perfectly represent the populations from which they were drawn, you should expect some instability in the results obtained from each sample. This instability is usually

referred to as "sampling error." The presence of sampling error is what makes drawing inferences about populations from samples difficult. One of the key goals of statistical theory is to estimate the amount of sampling error that is likely to be present in different statistical procedures and tests, and thereby gaining some idea about the amount of risk involved in using a particular procedure.

Statistical significance tests can be thought of as decision aids. That is, these tests can help you reach conclusions about whether the findings of your particular study are likely to represent real population effects, or whether they fall within the range of outcomes that might be produced by random sampling error. For example, there are two possible interpretations of the findings in this study of reading instruction:

1 - the difference between average scores from the two programs is so small that it might reasonably represent nothing more than sampling error

vs.

2 - the difference between average scores from the two programs is so large that it cannot be reasonably explained in terms of sampling error

The most common statistical procedure in the social and behavioral sciences is to pit a null hypothesis (H_0) against an alternative (H_1). In this example, the null and alternative hypotheses might take the forms:

H_0 - Reading instruction has no effect. It doesn't matter how you teach children to read because in the population, there is no difference in the average scores of children receiving either method of instruction

vs.

H_1 - Reading instruction has an effect. It does matter how you teach children to read because in the population; there is a difference in the average scores of children receiving different methods of instruction

Although null hypotheses usually refer to "no difference" or "no effect", it is important to understand that there is nothing magic about the hypothesis that the difference between two groups is zero. It might be perfectly reasonable to evaluate the following set of possibilities:

H_0 - In the population, the difference in the average scores of those receiving these two methods of reading instruction is six points

vs.

H_1- In the population, the difference in the average scores of those receiving these two methods of reading instruction is not six points

Another possible set of hypotheses is:

H_0 – In the population, the new method of reading instruction is not better than the old method; the new method might even be worse

vs.

H_1- In the population, the new method of reading instruction is better than the old method

This set of hypotheses leads to what is often called a “one-tailed” statistical test, in which the researcher not only asserts that there is a real difference between these two methods, but also describes the direction or the nature of this difference (i.e., that the new method is not just different from the old one, it is also better). We will discuss

one-tailed tests in several sections of this book, but in most cases will focus on the more widely-used two tailed tests that compare the null hypothesis that nothing happened with the alternative hypothesis that something happened; unless we specifically note otherwise, the traditional null hypothesis tests discussed in this book will be assumed to be two-tailed. However, the minimum effect tests we introduce in Chapter 2 and discuss extensively throughout the book have all of the advantages and few of the drawbacks of traditional one-tailed tests of the null hypothesis.

Beginning of Boxed section

Null Hypotheses vs. Nil Hypotheses

The most common structure for tests of statistical significance is to pit the null hypothesis that treatments have no effect, or that there is no difference between groups, or that there is no correlation between two variables against the alternative hypotheses that there is some treatment effect. In fact, this structure is so common that most people assume that the "null hypothesis" is essentially a statement that there is no difference between groups, no treatment effect, no correlation between variables, etc. This is not true. The null hypothesis is simply the hypothesis you actually test, and if you reject the null, you are left with the alternative. That is, if you reject the hypothesis that the effect of an intervention of treatment is "X", you are left to conclude that the alternative hypotheses that the effect of treatments is "not-X" must be true. If you test and reject the hypothesis that treatments have no effect, you are left with the conclusion that they must have some effect. If you test and reject the hypothesis that a particular diet will lead to a 20% weight loss, you are left with the conclusion that the diet will not lead to a 20% weight loss (if might have no effect, if might have a smaller effect, it might even have a larger effect).

Following Cohen's (1994) suggestion, we think it is useful to distinguish between the null hypothesis in general and its very special and very common form, the "nil hypothesis" – i.e., the hypothesis that treatments, interventions, etc. have no effect

whatsoever. The nil hypothesis is common because it is very easy to test and because it leaves you with a fairly simple and concrete alternative. If you reject the nil hypothesis that nothing happened, the alternative hypothesis you should accept is that something happened. However, as we will show in this chapter and in the chapters that follow, there are often important advantages to testing null hypotheses that are broader than the traditional nil hypothesis.

End of Boxed section

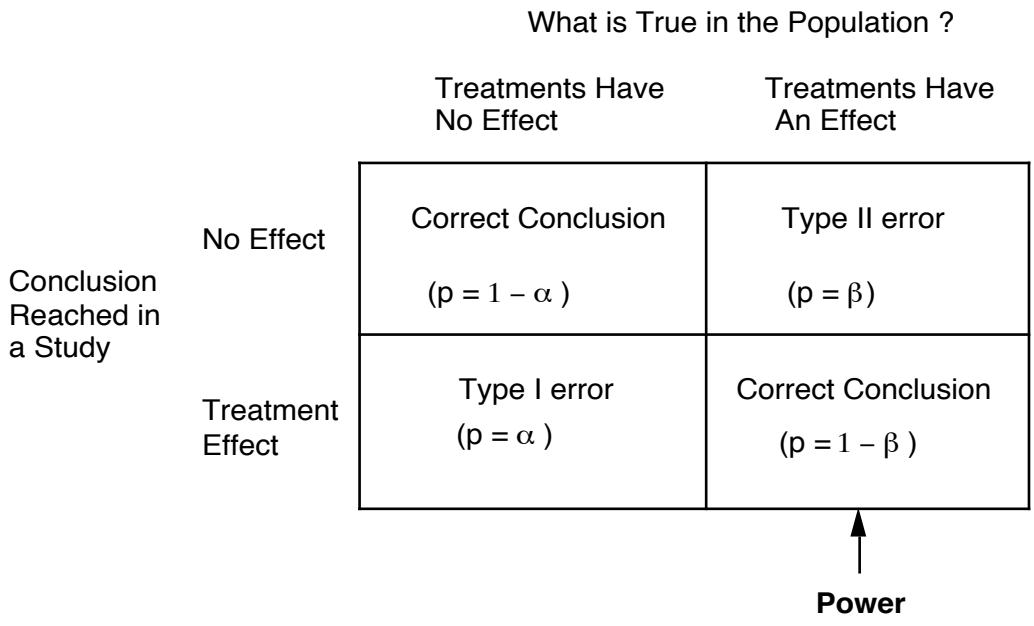
Most treatments of power analysis focus on the statistical power of tests of the nil hypothesis – i.e., tests of the hypothesis that treatments or interventions have no effect whatsoever. However, there are a number of advantages to posing and testing substantive hypotheses about the size of treatment effects (Murphy & Myors, 1999). For example, it is easy to test the hypothesis that the effects of treatments are negligibly small (e.g., they account for 1% or less of the variance in outcomes, or that the standardized mean difference is .10 or less). If you test and reject this hypothesis, you are left with the alternative hypothesis that the effect of treatments is not negligibly small, but rather large enough to deserve at least some attention. The methods of power analysis described in this book are easily extended to such minimum-effect tests, and are not limited to traditional tests of the null hypothesis that treatments have no effect.

What determines the outcomes of statistical tests? There are four outcomes that are possible when you use the results obtained in a particular sample to draw inferences about a population; these outcomes are shown in Figure 1-1.

Insert Figure 1-1 about here

As Figure 1-1 shows, there are two ways to make errors when testing hypotheses. First, it is possible that the treatment (e.g., new method of instruction) has no real effect in the population, but the results in your sample might lead you to believe

Figure 1-1
Outcomes of Statistical Tests



that it does have some effect. If the results of this study lead you to incorrectly conclude that the new method of instruction does work better than the current method, when in fact there were no differences, you would be making a *Type I* error (sometimes called an *alpha* error). Type I errors might lead you to waste time and resources by pursuing what are essentially dead ends, and researchers have traditionally gone to great lengths to avoid Type I errors.

There is an extensive literature dealing with methods of estimating and minimizing the occurrence of Type I errors (e.g., Keselman, Miller & Holland, 2011; Zwick & Marascuilo, 1984). The probability of making a Type I error is in part a function

of the standard or decision criterion used in testing your hypothesis (often referred to as alpha, or α). A very lenient standard (e.g., if there is any difference between the two samples, you will conclude that there is also a difference in the population) might lead to more frequent Type I errors, whereas a more stringent standard might lead to fewer Type I errors. [1]

A second type of error (referred to as *Type II* error, or a *beta* error) is also common in statistical hypothesis testing (Cohen, 1994; Sedlmeier & Gigerenzer, 1989). A Type II error occurs when you conclude in favor of H_0, when in fact H_1 is true. For example, if you conclude that there are no real differences in the outcomes of these two methods of instruction, when in fact one really is better than the other in the population, you have made a Type II error.

Statistical power analysis is concerned with Type II errors. i (i.e., if the probability of making a Type II error is β, power = 1 - β). Another way of saying this is to note that power is the (conditional probability) probability that you will *avoid* a Type II error. Studies with high levels of statistical power will rarely fail to detect the effects of treatments. If we assume that most treatments have at least *some* effect, the statistical power of a study often translates into the probability that the study will lead to the correct conclusion - i.e, that it will detect the effects of treatments.

Beginning of Boxed section

Understanding Conditional Probability

Conditional probability if an important concept in power analysis, and it can be a difficult on to understand. An example might help. Suppose a doctor sees 100 patients, and five of them have prostate cancer. You bump into one of this doctor's patients in the street. The simple probability that this particular patient has prostate cancer is:

Probability (cancer) = # who have cancer/total # of patients = 5/100 = .05

Assessments of conditional probability always require some additional information. Suppose that all patients are given routine screening exams for prostate cancer, and 20 patients receive high scores (indicating higher cancer risk). Suppose further, that 4 of these 20 patients actually have cancer. Assessments of conditional probability ask what is the likelihood of cancer given that we know a person has received a high score on the cancer screening test. There are only 20 people like this, and an assessment of the conditional probability of cancer, given a high score on this test is defined by

Probability(cancer|high score on screening test) = 4/20 =.20

In diagnostic testing, this type of conditional probability is referred to as the sensitivity of the test. The analysis above suggests that the test is not very sensitive. 16 of the 20 people who will get a cancer scare on the basis of the test will not in fact have cancer. Sensitivity can be contrasted with specificity, which is defined by a comparison between those who do not have cancer with those who do not receive high scores on the screening test. Sensitivity is also a type of conditional probability, defined here as

Probability(no cancer|low score on screening test) = 79/80 = .98

In this context, the test is very specific. People who receive a low score on the test almost never have cancer. The best diagnostic tests have high sensitivity (i.e., they almost always detect the condition) and high specificity (i.e., they rarely give healthy patients a false cancer scare).

The key difference between simple and conditional probability is the idea of a condition. That is, specificity is the likelihood that you are cancer free given that you receive a low test score. The idea of a condition also illustrates why we have such a low opinion of traditional null hypothesis testing, which evaluates the probability that (for example) two sample means will be found to be significantly different given the assumption that there is no difference in the population. We think there are ample reasons to question the belief that there are absolutely no differences between the populations from which just about any two samples are drawn, which means that the central starting point of traditional null hypothesis is that it depends on a condition that almost never really applies.

End of Boxed section

Effects of sensitivity, effect size, and decision criteria on power. The power of a statistical test is a function of its sensitivity, the size of the effect in the population, and the standards or criteria used to test statistical hypotheses. Tests have higher levels of statistical power when:

1. Studies are highly sensitive. Researchers can increase sensitivity by using better measures, or using study designs that allow them to control for unwanted sources of variability in your data (for the moment, we will define sensitivity in terms of the degree to which sampling error introduces imprecision into the results of a study; a fuller definition will be presented later in this chapter). The simplest method of increasing the sensitivity of a study is to increase its sample size (*N*) or the number of observations. [2] As *N* increases, statistical estimates become more precise and the power of statistical tests increase.

2. Effect sizes (ES) are large. Different treatments have different effects. It is easiest to detect the effect of a treatment if that effect is large (e.g., when treatment outcomes are very different, when treatments account for a substantial proportion of variance in outcomes; we will discuss specific measures of effect size later in this

chapter and in the chapters that follow). When treatments have very small effects, these effects can be difficult to reliably detect. As ES values increase, power increases.

3. Criteria for statistical significance are lenient. Researchers must make a decision about the standards that are required to reject H_0. It is easier to reject H_0 when the significance criterion, or alpha (α) level, is .05 than when it is .01 or .001. As the standard for determining significance becomes more lenient, power increases.

Power is highest when all three of these conditions are met (i.e., sensitive study, large effect, lenient criterion for rejecting the null hypothesis). In practice, sample size (which affects sensitivity) is probably the more important determinant of power. Effect sizes in the social and behavioral sciences tend to be small or moderate (if the effect of a treatment is so large that it can be seen by the naked eye, even in small samples, there may be little reason to test for it statistically), and researchers are often unwilling to abandon the traditional criteria for statistical significance that are accepted in their field (usually, alpha levels of .05 or .01; Cowles & Davis, 1982). Thus, effect sizes and decision criteria tend to be similar across a wide range of studies. In contrast, sample sizes vary considerably, and they directly impact levels of power.

With a sufficiently large *N*, virtually any test statistic will be "significantly" different from zero, and virtually any nil hypothesis can be rejected. Large *N* makes statistical tests highly sensitive, and virtually any specific point hypothesis (e.g., the difference between two treatments is zero, the difference between two reading programs is six points) can be rejected if the study is sufficiently sensitive. For example, suppose you are testing a new medicine that will result in a .0000001% increase in success rates for treating cancer. This increase is larger than zero, and if researchers include enough subjects in a study evaluating this treatment, they will almost certainly conclude that the new treatment is statistically different from existing treatments. On the other hand if very small samples are used to evaluate a treatment that has a real and substantial effect,

statistical power might be so low that you incorrectly conclude that the new treatment is not different from existing treatments.

Studies can have very low levels of power (i.e., are likely to make Type II errors) when they use small samples, when the effect being studied is a small one, and/or when stringent criteria are used to define a "significant" result. The worst case occurs when a researcher uses a small sample to study a treatment that has a very small effect, *and* he or she uses a very strict standard for rejecting the null hypothesis. Under those conditions, Type II errors may be the norm. To put it simply, studies that use small samples and stringent criteria for statistical significance to examine treatments that have small effects will almost always lead to the wrong conclusion about those treatments (i.e., to the conclusion that treatments have no effect whatsoever).

The Mechanics of Power Analysis

When a sample is drawn from a population, the exact value of any statistic (e.g., the mean, difference between two group means) is uncertain, and that uncertainty is reflected by a statistical distribution. Suppose, for example, that you evaluate a treatment that you expect has no real effect (e.g., you use astrology to advise people about career choices) by comparing outcomes in groups who receive this treatment to outcomes in groups who do not receive it (control groups). You will not always find that treatment and control groups have exactly the same scores, even if the treatment has no real effect. Rather, there is some range of values can be expected for any test statistic in a study like this, and the standards used to determine statistical significance are based on this range or distribution of values. In traditional null hypothesis testing, a test statistic is "statistically significant" at the .05 level if its actual value is outside of the range of values you would observe 95% of the time in studies where the treatment had no real effect. If the test statistic is outside of this range, the usually inference is that the treatment *did* have some real effect.

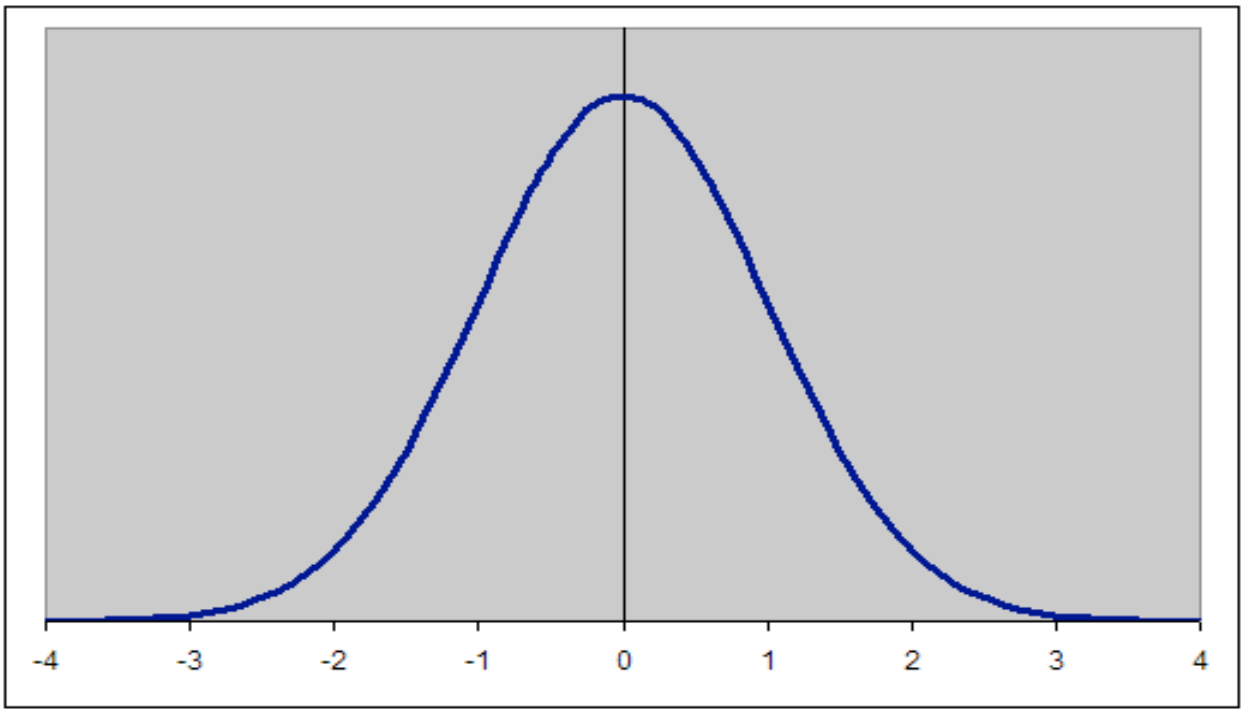
For example, suppose that 62 people are randomly assigned to treatment and control groups, and the *t* statistic is used to compare the means of the two groups. If the treatment has no effect whatsoever, the *t* statistic should usually be near zero, and will have a value less than or equal to approximately 2.00 in 95% of the time. If the *t* statistic obtained in a study is larger than 2.00, you can safely infer that treatments are very likely to have some effect; if there was no real treatment effect, values above 2.00 would be a very rare event.

Beginning of Boxed Section

Understanding Sampling Distributions

In the example above, 62 people are randomly assigned to groups that either receive astrology-based career advice or who do not receive such advice. Even though you might expect that the treatment has no real effect, you would probably not expect that the difference between the average level of career success of these two groups will always be exactly zero. Sometimes the astrology group might do better and sometimes it might do worse.

Suppose you repeated this experiment 1,000 times, and noted the difference between the average level of career success in the two groups. The distribution of scores would look something like the figure below:



This distribution is referred to as a sampling distribution, and it illustrates the extent to which differences between the means of these two groups might be expected to vary as a result of chance or sampling error. Most of the time, the differences between these groups should be near zero, because we expect that advice based on astrology has no real systematic effect. The variance of this distribution illustrates the range of differences in outcomes you might expect if your hypothesis that astrology has no real effect. In this case, about 95% of the time, you expect the difference between the astrology and the no-astrology groups to be about two points or less. If you find a bigger difference between groups (suppose the average success score for the astrology group is five points higher than the average for the no-astrology group), you should reject the null hypothesis that the career advice has no systematic effect.

You might ask why anyone in their right mind would repeat this study 1,000 times. Luckily, statistical theory allows us to estimate sampling distributions on the basis of a few simple statistics. Virtually all of the statistical tests discussed in this book are conducted by comparing the value of some test statistic to its sampling distribution,

so understanding the idea of a sampling distribution is essential to understanding hypothesis testing and statistical power.

End of Boxed Section

As the example above suggests, if treatments have no effect whatsoever in the population, you should not expect to always find a difference of precisely zero between samples of those who receive the treatment and those who do not. Rather, there is some range of values that might be found for any test statistic in a sample (e.g., in the example cited earlier, you expect the value of the difference in the two means to be near zero, but you also know it might range from about -2.00 to +2.00). The same is true if treatments have a real effect. For example, if a researcher expects that the mean in a treatment group that receives career advice based on valid measures of work interests will be 10 points higher than the mean in a control group (e.g., because this is the size of the difference in the population), that researcher should also expect some variability around that figure. Sometimes, the difference between two samples might be 9 points, and sometimes it might be 11 or 12. The key to power analysis is estimating the range of values one might reasonably expect for some test statistic if the real effect of treatments is small, or medium, or large.

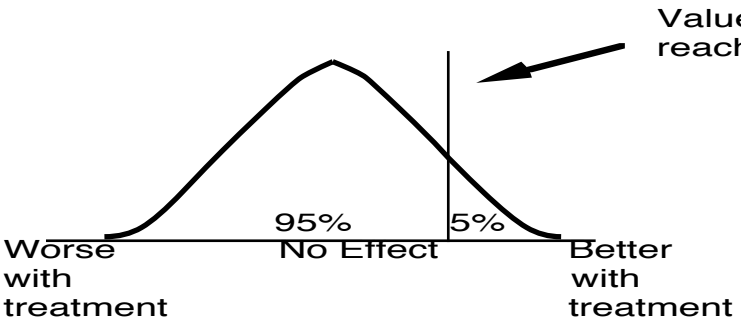
Figure 1-2 illustrates the key ideas in statistical power analysis.

Insert Figure 1-2 about here

Suppose you use a t-test to determine whether the difference in average reading test scores of 3,000 pupils randomly assigned to two different types of reading instruction are statistically significant. You do not make any specific prediction about which reading program will be better, and therefore test the two-tailed hypothesis that the two programs lead to systematically different outcomes. To be "statistically significant", the value of this test statistic must be 1.96 or larger. As Figure 1-2 suggests, the likelihood

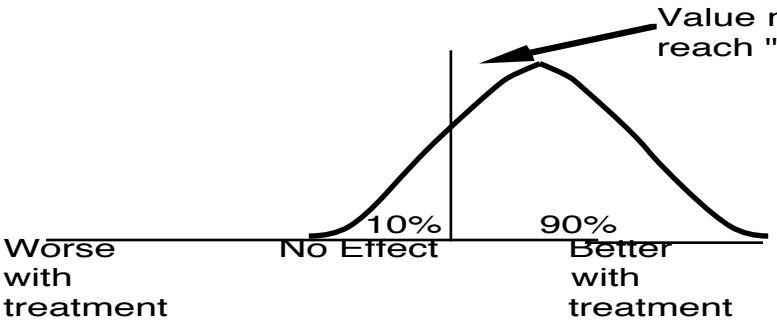
Figure 1-2 Essentials of Power Analysis

1. Distribution Expected if Treatments Have No Effect *



2. Distribution Expected if Treatments Have a Small Effect *

3. Distribution Expected if Treatments Have a Large Effect *



* - depending on the test statistic in question, the distributions might take different forms, but the essential features of this figure would apply to any test statistic

you will reject the null hypothesis that there is no difference between the two groups depends substantially on whether the true effect of treatments is small or large.

If the null hypothesis that there is no real effect was true, you would expect to find values of 1.96 or higher for this test statistic in 5 tests out of every 100 performed (i.e., α = .05). This is illustrated in Section 1 of Figure 1-2. Section 2 of Figure 1-2 illustrates the distribution of test statistic values you might expect if treatments had a small effect on the dependent variable. You might notice that the distribution of test statistics you would expect to find in studies of a treatment with this sort of effect has shifted a bit, and that in this case 25% of the values you might expect to find are greater than or equal to 1.96. That is, if you run a study under the scenario illustrated in Section 2 of this figure (i.e., treatments have a small effect), the probability you will reject the null hypothesis is .25. Section 3 of Figure 1-2 illustrates the distribution of values you might expect if the true effect of treatments is large. In this distribution, 90% of the values are 2.00 or greater, and the probability you will reject the null hypothesis is .90. ***The power of a statistical test is the proportion of the distribution of test statistics expected for a study like this that is above the critical value used to establish statistical significance.***

The qualifier “for a study like this” is important, because the distribution of test statistics you should reasonably expect in a particular study depends on both the population effect size and the sample size. If the power of a study is .80, that is the same thing as saying that if you draw a distribution of the test statistic values you expect to find based on the population effect size and the sample size, 80% of these will be equal to or larger than the critical value needed to reject the null hypothesis.

No matter what hypothesis you are testing, or what statistic you are using to test that hypothesis, power analysis always involves three basic steps, which are listed in Table 1-1. First, a criterion or critical value for "statistical significance" must be

---
Insert Table 1-1 about here
---

established. For example, the tables found in the back of virtually any statistics textbook can be used to determine such critical values for testing the traditional null hypothesis. If the test statistic a researcher computes exceeds this critical value, the researcher will reject the null hypothesis. However, these tables are not the only basis for setting such a criterion. Suppose you wanted to test the hypothesis that the effects of treatments are so small that they can safely be ignored. This might involve specifying some range of effects that would be designated as "negligible", and then determining the critical value of a statistic needed to reject this hypothesis. Chapter 2, shows how such tests are performed, and lays out the implications of such hypothesis testing strategies for statistical power analysis

Power analysis required researchers to must make their best guess of the size of effect treatments are likely to have on the dependent variable(s); methods of estimating effect sizes are discussed later in this chapter. As we noted earlier, if there are good reasons to believe that treatments have a very large effect, it should be quite easy to reject the null hypothesis. On the other hand, if the true effects of treatments are small and subtle, it might be very hard to reject the hypothesis that treatments have no real effect.

Once you have estimated ES, it is also possible to use that estimate to describe the distribution of test statistics that should be expected. We will describe this process in more detail in Chapter 2, but a simple example will show you what we mean.

Table 1-1

The Three Steps to Determining Statistical Power

1. Establish a criterion or critical value for statistical significance
    * what is the hypothesis that is being tested? (e.g., traditional null hypothesis, minimum-effect tests)
    * what level of confidence is desired? (e.g., a = .05 vs. a = .01)
    * what is the critical value for your test statistic [these critical values are determined on the basis of the degrees of freedom (df) for the test and the desired confidence level level)
2. Estimate the effect size (ES)
    * are treatments expected to have large, medium, or small effects?
    * what is the range of values researchers expect to find for the test statistic, given this ES?

3. Determine where the critical value lies in relationship to the distribution of test statistics expected if the null hypothesis is true (i.e., the sampling distribution)

The power of a statistical test is the proportion of the distribution of test statistics expected for a study (based on the sample size and the estimated ES) that is above the critical value used to establish statistical significance

Suppose you are using a *t*- test to assess the difference in the mean scores of those receiving two different treatments. If there was no real difference between the treatments, you would expect to find *t* values near zero most of the time, and you can use statistical theory to tell how much these *t* values might depart from zero as a result of sampling error. The *t* tables in most statistics textbooks tell you how much variability you might expect with samples of different sizes, and once the mean (here, zero) and the standard deviation of this distribution are known, it is easy to estimate what proportion of the distribution falls above or below any critical value. If there is a large difference between the treatments (e.g., the dependent variable has a mean of 500 and a standard deviation of 100, and the mean for one treatment is usually 80 points higher than the mean for another), large *t* values should be expected most of the time.

The final step in power analysis is a comparison between the values obtained in the first two steps. For example, if you determine that a *t* value of 1.96 is needed to reject the null hypothesis, and also determine that because the treatments being studied have very large effects you are likely to find *t* values of 1.96 or greater 90% of the time, the power of this test - i.e., power is .90.

Sensitivity and power. Sensitivity refers to the precision with which a statistical test distinguishes between true treatment effects and differences in scores that are the result of sampling error. As noted above, the sensitivity of statistical tests is largely a function of the sample size. Large samples provide very precise estimates of population parameters, whereas small samples produce results than can be unstable and untrustworthy. For example, if 6 children in 10 do better with a new reading curriculum than with the old one, this might reflect nothing more than simple sampling error. If 600 out of 1,000 children do better with the new curriculum, this is powerful and convincing evidence that there are real differences between the new curriculum and the old one.

In a study with low sensitivity, there is considerable uncertainty about statistical outcomes. As a result, it might be possible to find a large treatment effect in a sample, even though there is no true treatment effect in the population. This translates into both substantial variability in study outcomes and the need for relatively demanding tests of "statistical significance". If outcomes can vary substantially from study to study, researchers need a to observe relatively large effects to be confident that they represents true treatment effects and not mere sampling error. As a result, it is often difficult to reject the hypothesis that there is no true effect when small samples are used, and many Type II errors should be expected.

In a highly sensitive study, there is very little uncertainty or random variation in study outcomes, and virtually any difference between treatment and control groups is likely to be accepted as an indication that the treatment has an effect in the population.

Effect size and power. Effect size is a key concept in statistical power analysis (Cohen, 1988; Rosenthal, 1991; Tatsuoka, 1993a). At the simplest level, effect size measures provide an index of how much impact treatments actually have on the dependent variable; if H_0 states that treatments have no impact whatsoever, the effect size can be though of as an index of just how wrong the null hypothesis is.

One of the most common ES measures is the standardized mean difference, *d*, defined as $d = (M_t - M_c)/SD$, where M_t and M_c are the treatment and control group means, respectively, and SD is the pooled standard deviation. By expressing the difference in group means in standard deviation units, the *d* statistic provides a simple metric that allows one to compare treatment effects from different studies, areas or research, etc., without having to keep track of the units of measurement used in different studies or areas of research. For example, Lipsey and Wilson (1993) cataloged the effects of a wide range of psychological, educational, and behavioral treatments, all expressed in terms of *d*. Examples of interventions in these areas that

have relatively small, moderately large, and large effects on specific sets of outcomes are presented in Table 1-2.

Insert Table 1-2 about here

For example, worksite smoking cessation/reduction programs have a relatively small effect on quit rates (*d* = .21). The effects of class size on achievement or of juvenile delinquency programs on delinquency outcomes are similarly small. Concretely, a *d* value of .20 means that the difference between the average score of those who receive the treatment and those who do not is only 20% as large as the standard deviation of the outcome measure within each of the treatment groups. This standard deviation measures the variability in outcomes, independent of treatments, so *d* = .20 indicates that the average effect of treatments is only 1/5th as large as the variability in outcomes among people who receive the same treatments. In contrast, interventions such as psychotherapy, meditation and relaxation, or positive reinforcement in the classroom have relatively large effects on outcomes such as functioning levels, blood pressure, and learning (d values range from .85 to 1.17).

Table 1-2

Examples of Effect Sizes Reported in Lipsey and Wilson (1993) Review

	Dependent Variable	d
Small Effects (d = .20)		
Treatment programs for juvenile delinquents	delinquency outcomes	.17
Worksite smoking cessation/reduction programs	quit rates	.21
Small vs. large class size, all grade levels	achievement measures	.20
Medium Effects (d = .50)		
Behavior therapy vs. placebo controls	various outcomes	.51
Chronic disease patient education	compliance and health	.52
Enrichment programs for gifted children	cognitive, creativity, affective outcomes	.55
Large Effects (d = .80)		
Psychotherapy	various outcomes	.85
Meditation and relaxation techniques	blood pressure	.93
Positive reinforcement in the classroom	learning	1.17

Beginning of Boxed Section

d vs. delta vs. *g*

There are several effect size measures that represent variations on the same theme as the *d* statistic described here. *d* represents the difference between means, expressed in standard deviation units. Which standard deviation you should use to describe this difference, however, is not always a simple question. Glass introduced a statistic delta (Δ) which divides the difference in means by the standard deviation in the control group – the rationale being that treatments or interventions can change both the mean and the standard deviation (Glass, McGaw & Smith, 1981). Cohen's *d* uses the average of the SDs in each group as a standard for comparison. Hedges introduced a statistic *g*, to correct for biases in *d* when small samples are used. There are different *d* values that might be obtained, depending on whether the population standard deviation is known, or whether (as is usual) it is estimated on the basis of the sample standard deviations.

When samples are very small, these different statistics may yield different values, but as *N* increases, they tend to be more and more similar. More to the point, they all express the same concept, that the differences between means are best understood in comparison to a measure of the variability of scores – i.e., the standard deviation.

End of Boxed Section

It is important to keep in mind that "small", "medium" or "large" effect refers to the size of the effect, but not necessarily to its importance. For example, a new security screening procedure might lead to a small change in rates of detecting threats, but if this change translates into hundreds of lives saved at a small cost, the effect might be judged to be both important and worth paying attention to.

When the true treatment effect is very small, it might be hard to accurately and consistently detect this effect in successive samples. For example, aspirin can be

useful in reducing heart attacks, but the effects are relatively small (*d* =.068; See, however, Rosenthal, 1993). As a result, studies of 20 or 30 patients taking aspirin or a placebo will not consistently detect the true and life-saving effects of this drug. Large-sample studies, however, provide compelling evidence of the consistent effect of aspirin on heart attacks. On the other hand, if the effect is relatively large, it is easy to detect, even with a relatively small sample. For example, cognitive ability has a strong influence on performance in school (*d* is about 1.10), and the effects of individual differences in cognitive ability are readily noticeable even in small samples of students.

Decision criteria and power. Finally, the standard or decision criteria used in hypothesis testing has a critical impact on statistical power. The standards that are used to test statistical hypotheses are usually set with a goal of minimizing Type I errors; alpha levels are usually set at .05, .01, or some other similarly low level, reflecting a strong bias against treating study outcomes that might be due to nothing more than sampling error as meaningful (Cowles & Davis, 1982). Setting a more lenient standard makes it easier to reject the null hypothesis, and while this can lead to Type I errors in those rare cases where the null is actually true, anything that makes it easier to reject the null hypothesis also increases the statistical power of your study.

As Figure 1-1 shows, there is always a tradeoff between Type I and Type II errors. If you make it very difficult to reject the null hypothesis, you will minimize Type I errors (incorrect rejections), but you will also increase the number of Type II errors. That is, if you rarely reject the null, you will often incorrectly dismiss sample results as mere sampling error, when they may in fact indicate the true effects of treatments. Numerous authors have noted that procedures to control or minimize Type I errors can substantially reduce statistical power, and may cause more problems (i.e., Type II errors) than they solve (Cohen, 1994; Sedlmeier & Gigerenzer, 1989).

Power analysis and the general linear model. In the chapters that follow, we will describe a simple and general model for statistical power analysis. This model is based

on the widely-used *F* statistic. This statistic and variations on the *d* are used to test a wide range of statistical hypotheses in the context of the general linear model (Cohen & Cohen, 1983; Horton, 1978; Tatsuoka, 1993b). The general linear model provides the basis for correlation, multiple regression, analysis of variance, discriminant analysis, and all of the variations of these techniques. The general linear model subsumes a large proportion of the statistics that are widely used in the behavioral and social sciences, and by tying statistical power analysis this model, we will show how the same simple set of techniques can be applied to an extraordinary range of statistical analyses.

Statistical Power of Research in the Social and Behavioral Sciences

Research in the social and behavioral sciences often shows shockingly low levels of power. Starting with Cohen's (1962) review of research published in *Journal of Abnormal and Social Psychology*, studies in psychology, education, communication, journalism, and other related fields have routinely documented power in the range of .20 - .50 for detecting small to medium treatment effects (Sedlmeier & Gigerenzer, 1989). Despite decades of warnings about the consequences of low levels of statistical power in the behavioral and social sciences, the level of power encountered in published studies is lower than .50 (Mone, Mueller & Mauland, 1996). In other words, it is typical for studies in these areas to have less than a 50% chance of rejecting the null hypothesis. If you believe that the null hypothesis is virtually always wrong (i.e., that treatments have at least some effect, even if it is a very small one), this means that at least half of all studies in the social and behavioral sciences, (perhaps as many as 80%) are likely to reach the wrong conclusion by making a Type II error, when testing the null hypothesis.

These figures are even more startling and discouraging when you realize that these reviews have examined the statistical power of *published research*. Given the strong biases against publishing methodologically suspect studies or studies reporting null results, it is likely that the studies that survive the editorial review process are better

than the norm, that they show stronger effects than similar unpublished studies, and that the statistical power of unpublished studies is even lower than the power of published studies.

Studies that do not reject the null hypothesis are often regarded by researchers as failures. The levels of power reported above suggest that "failure", defined in these terms, is quite common. If a treatment effect is small, and a study is designed with a power level of .20 (which is depressingly common), researchers are four times as likely to fail (i.e., fail to reject the null) as to succeed. Power of .50 suggests that the outcome of a study is basically like the flip of a coin. A researcher whose study has power of .50 is just as likely to fail to reject the null hypothesis as he or she is to succeed. It is likely that much of the apparent inconsistency in research findings is due to nothing more than inadequate power (Schmidt, 1992). If 100 studies are conducted, each with power of .50, about half of them will reject the null and about half will not. Given the stark implications of low power, it is important to consider *why* research in the social and behavioral sciences is so often conducted in a way in which failure is more likely than success.

The most obvious explanation for low level of power in the social and behavioral sciences is the belief that social scientists tend to study treatments, interventions, etc. that have small and unreliable effects. Until recently, this explanation was widely accepted, but the widespread use of meta-analysis in integrating scientific literature suggests that this is not necessarily the case. There is now ample evidence from literally hundred of analyses of thousands of individual studies that the treatments, interventions, and the like studied by behavioral and social scientists have substantial and meaningful effects (Haase, Waechter & Solomon, 1982; Hunter & Hirsh, 1987; Lipsey, 1990; Lipsey & Wilson, 1993; Schmitt, Gooding, Noe & Kirsch, 1984); these effects are of a similar magnitude as many of the effects reported in the physical sciences (Hedges, 1987).

A second possibility is that the decision criteria used to define "statistical significance" are too stringent. We will argue in several of the chapters that follow that researchers are often too concerned with Type I errors and insufficiently concerned with statistical power. However, the use of overly stringent decision criteria is probably not the best explanation for low levels of statistical power.

The best explanation for the low levels of power observed in many areas of research is many studies use samples that are much too small to provide accurate and credible results. Researchers routinely use samples of 20, 50, or 75 observations to make inferences about population parameters. When sample results are unreliable, it is necessary to set some strict standard to distinguish real treatment effects from fluctuations in the data that are due to simple sampling error, and studies with these small samples often fail to reject null hypotheses, even when the population treatment effect is fairly large.

On the other hand, very large samples will allow you to reject the null hypothesis even when it is very nearly true - i.e., when the effect of treatments is very small. In fact, the effects of sample size on statistical power are so profound that it is tempting to conclude that a significance test is little more than a roundabout measure of how large the sample is. If the sample is sufficiently small, you will virtually never reject the null hypothesis. If the sample is sufficiently large, you will virtually always reject the null hypothesis.

Using Power Analysis

Statistical power analysis can be used for both planning and diagnosis. Power analysis is frequently used in designing research studies. The results of power analysis can help in determining how large your sample should be, or in deciding what criterion should be used to define "statistical significance". Power analysis can also be used as a diagnostic tool, to determine whether a specific study has adequate power for specific purposes, or to identify the sort of effects that can be reliably detected in that study.

Because power is a function of the sensitivity of your study (which is essentially a function of *N*), the size of the effect in the population (ES), and the decision criterion that is used to determine statistical significance, we can solve for any of the four values (i.e, power, *N*, ES, α), given the other three. However, none of these values is necessarily known in advance, although some values may be set by convention. The criterion for statistical significance (i.e., α) is often set at .05 or .01 by convention, but there is nothing sacred about these values. As we will note later, one important use of power analysis is in making decisions about what criteria should be used to describe a result as "significant".

Boxed Section

The Meaning of Statistical Significance

Suppose a study leads to the conclusion that "there is a statistically significant correlation between the personality trait of conscientiousness and job performance". What does "statistically significant" mean?

"Statistically significant" clearly does not mean that this correlation is large, meaningful, or important (although it might be all of these). If the sample size is large, a correlation that is quite small will still be "statistically significant". For example, if *N* = 20,000, a correlation of r = .02 will be significantly (α=.05) different from zero. The term "statistically significant" can be thought of as shorthand for the following statement:

"In this particular study, there is sufficient evidence to allow the researcher to reliably distinguish (with a level of confidence defined by the alpha level) between the observed correlation of .02 and a correlation of zero"

In other words, the term "statistically significant" does not describe the result of a study, but rather describes the sort of result this particular study can reliably detect. The same correlation will be statistically significant in some studies (e.g., those that use a large *N* or a lenient alpha) and not significant in others. In the end, "statistically significant" usually says more about the design of the study than about the results.

Studies that are designed with high levels of statistical power will, by definition, usually produce significant results. Studies that are designed with low levels of power will not yield significant results. A significant test usually tells you more about the study design than about the substantive phenomenon being studied.

Boxed Section

The ES depends on the treatment, phenomenon, or variable you are studying, and is usually not known in advance. Sample size is rarely set in advance, and *N* often depends on some combination of luck and resources on the part of the investigator. Actual power levels are rarely known, and it can be difficult to obtain sensible advice about how much power you should have. It is important to understand how each of the parameters involved is determined when conducting a power analysis.

Determining the effect size. There is a built-in dilemma in power analysis. In order to determine the statistical power of a study, ES must be known. But if you already knew the exact strength of the effect the particular treatment, intervention, etc., you would not need to do the study! The whole point of doing a study is to find out what effect the treatment has, and the true ES in the population is unlikely to ever be known.

Statistical power analyses are always based on *estimates* of ES. In many areas of study, there is a substantial body of theory and empirical research that will provide a well-grounded estimate of ES. For example, there are literally hundreds of studies of the validity of cognitive ability tests as predictors of job performance (Hunter & Hirsch, 1987; Schmidt, 1992), and this literature suggests that the relationship between test scores and performance is consistently strong (corrected correlations of about .50 are common). Even where there is not an extensive literature available, researchers can often use their experience with similar studies to realistically estimate effect sizes.

When there is no good basis for estimating effect sizes, power analyses can still be carried out, by making a conservative estimate. A study that has adequate power to reliably detect small effects (e.g., a *d* of .20, or a correlation of .10) will also have adequate power to detect larger effects. On the other hand, if researchers design studies with the assumption that effects will be large, they might have insufficient power to detect small but important effects. Earlier, we noted that the effects of taking aspirin on heart attacks are relatively small, but that there is still a substantial payoff for taking the drug. If the initial research that led to the use of aspirin for this purpose had been conducted using small samples, the researchers would have had little chance of detecting the life-saving effect of aspirin.

Determining the desired level of power. In determining desired levels of power, researchers must weigh the risks of running studies without adequate power against the resources needed to attain high levels of power. Researchers can always achieve high levels of power by using very large samples, but the time and expense required may not always justify the effort.

There are no hard-and-fast rules about how much power is enough, but there does seem to be consensus about two things. First, if at all possible, power should be above .50. When power drops below .50, a study is more likely to fail (i.e., it is unlikely to reject the null hypothesis) than succeed. It is hard to justify designing studies in which failure is the most likely outcome. Second, power of .80 or above is usually judged to be adequate. The .80 convention is arbitrary (in the same way that significance criteria of .05 or .01 are arbitrary), but it seems to be widely accepted, and it can be rationally defended.

Power of .80 means that success (rejecting the null) is four times as likely as failure. It can be argued that some number other than four might represent a more acceptable level of risk (e.g., if power = .90, success is nine times as likely as failure), but it is often prohibitively difficult to achieve power much in excess of .80. For

example, to have a power of .80 in detecting a small treatment effect (where the difference between treatment and control groups is *d* = .20), a sample of about 775 subjects is needed. If power of .95 is desired, a sample of about 1300 subjects will be needed. Most power analyses specify .80 as the desired level of power to be achieved, and this convention seems to be widely accepted.

Applying power analysis. There are four ways to use power analysis: (1) in determining the sample size needed to achieve desired levels of power, (2) in determining the level of power in a study that is planned or has already been conducted, (3) in determining the size of effect that can be reliably detected by a particular study, and (4) in determining sensible criteria for "statistical significance". The chapters that follow will lay out the actual steps in doing a power analysis, but it is useful at this point to get a preview of the four potential applications of this method. Power analysis can be used in:

1. Determining sample size - given a particular ES, significance criterion and a desired level of power, it is easy to solve for the sample size needed. For example, if researchers think the correlation between a new test and performance on the job is .30, and they want to have at least an 80% chance of rejecting the null hypothesis (with a significance criterion of .05), they need a sample of about 80 cases. When planning a study, researchers should routinely use power analysis to help make sensible decisions about the number of subjects needed.

2. Determining power levels - if *N*, ES, and the criterion for statistical significance are known, researchers can use power analysis to determine the level of power for that study. For example, if the difference between treatment and control groups is small (e.g.,*d* = .20), there are 50 subjects in each group, and the significance criterion is α = .01, power will be only .05! Researchers should certainly expect that this study will fail to reject the null, and they might decide to change the design of this study considerably (e.g., use larger samples, more lenient criteria)

3. Determine ES levels - researchers can also determine what sort of effect could be reliably detected, given *N*, the desired level of power, and α. In the example above, a study with 50 subjects in both the treatment and control groups would have power of .80 to detect a very large effect (approximately *d* = .65) with a .01 significance criterion, or a large effect (*d* =.50) with a .05 significance criterion.

4. Determine criteria for statistical significance - given a specific effect, sample size, and power level, it is possible to determine the significance criterion. For example, if you expect a correlation coefficient to be .30, *N* = 67, and you want power to equal or exceed .80, you will need to use a significance criterion of α = .10 rather than the more common .05 or .01.

Hypothesis Tests vs. Confidence Intervals

Null hypothesis testing has been criticized on a number of grounds (e.g., Schmidt, 1996), but perhaps the most persuasive critique is that null hypothesis tests provide so little information. It is widely recognized that the use of confidence intervals and other methods of portraying levels of uncertainty about the outcomes of statistical procedures have many advantages over simple null hypothesis tests (Wilkinson et al., 1999).

Suppose a study is performed that examines the correlation between scores on an ability test and measures of performance in training. The authors find a correlation of *r* = .30, and on the basis of a null hypothesis test, decide that this value is significantly (e.g., at the .05 level) different from zero. That test tells them something, but it does not really tell them whether the finding that *r* = .30 represents a good or a poor estimate of the relationship between ability and training performance. A confidence interval would provide that sort of information.

Staying with this example, suppose researchers estimate the amount of variability expected in correlations from studies like this, and conclude that a 95% confidence interval ranges from .05 to .55. This confidence interval would tell

researchers exactly what they learned from the significance test - i.e., that they could be quite sure the correlation between ability and training performance was not zero. A confidence interval would also tell them that *r* = .30 might not be a good estimate of the correlation between ability and performance; the confidence interval suggests that this correlation could be much larger or much smaller than .30. Another researcher doing a similar study using a larger sample might find a much smaller confidence interval, indicating a good deal more certainty about the generalizability of sample results.

As the previous paragraph implies, most of the statements that can be made about statistical power also apply to confidence intervals. That is, if you design a study with low power, you will also find that it produces wide confidence intervals (i.e., that there is considerable uncertainty about the meaning of sample results). If you design studies to be sensitive and powerful, these studies will yield smaller confidence intervals. Although the focus of this book is on hypothesis tests, it is important to keep in mind that the same facets of the research design (*N*, the alpha level) that cause power to go up or down also cause confidence intervals to shrink or grow. A powerful study will not always yield precise results (e.g., power can be high in a poorly-designed study that examines a treatment that has very strong effects) but in most instances, whatever researchers do to increase power will also lead to smaller confidence intervals and to more precision in sample statistics.

By almost any criterion, statistical analyses that include confidence intervals are preferable to analyses that rely on null hypothesis tests alone. Unfortunately, the process of changing researchers' habits with regard to data analysis is very slow and uncertain, and too few researchers incorporate confidence intervals when they report the results of statistical analyses (Fidler, Thomason, Cumming, Finch & Leeman, 2004).

Beginning of Boxed Section

Accuracy in Parameter Estimation

Statistical power analysis is concerned to do with null hypothesis testing, and with the right null hypothesis (e.g., one that specifies a range of values rather than a single point) this approach can provide some real value. In particular, it can help you make good choices regarding sample size. An alternative perspective for determining the sample sizes needed in a particular study is to do so in terms of the desired level of accuracy in parameter estimation (AIPE). For example, if the focus of your study is on the squared multiple correlation between academic achievement and five measures of school quality, you could use power analysis to determine the sample size needed to reject the traditional null hypothesis that this correlation is exactly zero, or the sample size needed to reject the null hypothesis that this correlation is so small as to be trivial in its magnitude. The AIPE approach focuses on the level of accuracy desired in estimating the relationship between school quality and academic achievement. For example, if you believed that school quality accounted for 20% of the variance in achievement and you wanted to achieve a confidence interval sufficiently small to be relatively certain that the population R squared was between .10 and .30 in value, equations developed by Kelly (2008) show that you would need a sample of N=239.

Kelly and his colleagues have developed methods and software for determining the sample size needed to reach desired levels of accuracy in evaluating the standardized difference between two means (d - See Kelly & Rausch, 2006), squared multiple correlations (Kelly, 2008), and various parameters in structural equation modeling (Kelly & Lai, 2011; Lai & Kelly, 2012; Maxwell, Kelly & Rausch, 2008). As with power analysis, the biggest problem with the AIPE approach is that it requires you to have at least a general idea of the expected size of the effect in the population. As in power analysis, the best strategy for planning sample sizes when there is not a good

estimate of the effect size in the population, the best thing to do is to plan as if the population effect is a small one.

End of Boxed Section

What Can You Learn from a Null Hypothesis Test?

Tests of the traditional nil hypothesis, that the difference between two treatments in exactly zero can be informative, but unless they are accompanied with some sort of effect size measures and unless the power of these tests is known, it can be difficult to draw any sensible conclusions from this type of test. Consider the example of a study that compares two alternative treatments for colon cancer (e.g., radiation vs. chemotherapy). The study fails to reject the null hypothesis. If you do not know anything about the power of this test or about the likely size of the difference in the effects of these two treatments in the population, you are left with two alternatives.

1. The study has too little power to detect a real difference between these two treatments
2. The difference between these treatments is either exactly zero (something that is extremely unlikely) or it is so small that it is probably not meaningful

In this book, we are skeptical about the value of most null hypothesis tests because of the difficulty in determining which alternative of more likely to be true. In particular, we will show you why the failure to reject the null hypothesis in most studies is more likely to say something about the design of the study (i.e., the study did not have sufficient power) than about the phenomenon being studied. The methods of framing and testing minimum effect hypotheses introduced in Chapter 2 can be applied to reframe the questions that are asked in a null hypothesis test in a way that will make the outcomes of significance tests both more informative and more easily understood.

Summary

Power is defined as the probability that a study will reject the null hypothesis when it is in fact false. Studies with high statistical power are very likely to detect the effects of treatments, interventions, etc., whereas studies with low power can lead researchers to dismiss potentially important effects as sampling error. The statistical power of a test is a function of the size of the treatment effect in the population, the sample size, and the particular criteria used to define statistical significance. Although most discussions of power analysis are phrased in terms of traditional null hypothesis testing, where the hypothesis that treatments have no impact whatsoever is tested, power analysis can be fruitfully applied to any method of statistical hypothesis testing.

Statistical power analysis has received less attention in the behavioral and social sciences than it deserves. It is still routine in many areas for researchers to run studies with disastrously low levels of power. Statistical power analysis can and should be used to determine the number of subjects that should be included in a study, to estimate the likelihood that a study will reject the null hypothesis, to determine what sorts of effects can be reliably detected in a study, or to make rational decisions about the standards used to define "statistical significance". Each of these applications of power analysis is taken up in the chapters that follow.

One way to think about significance testing and power analysis is to think about a significance test as a way of determining whether or not something you think might be there really is there. For example, suppose you use a cheap telescope and look in the night sky to search for the moons of Jupiter. You probably will not see them. What should you conclude? We hope that after reading this chapter you will conclude that if you only had a more powerful telescope, you would see them. The same it true for differences between treatments, relationships between variables and the like. These differences and relationships might be quite small (they might be so small that they are

not important in any practical sense), but if you failed to detect them with a significance test, the conclusion you should reach is that a more powerful test might very well detect them. If you use the cheap telescope and fail to find the moons of Saturn, you will conclude that they must be gone, but will rather conclude that you need a better telescope. The same idea applied to tests of the traditional nil hypothesis. If you are going to do these tests at all, you should first be sure you have sufficient statistical power to make them useful.

Footnotes

1. It is important to note that Type I errors can only occur when the null hypothesis is actually true. If the null hypothesis is that there is no true treatment effect (a nil hypothesis) this will rarely be the case. As a result, Type I errors are probably quite rare in published tests of the traditional null hypothesis, and efforts to control these errors at the expense of making more Type II errors might be ill advised (Murphy, 1990).
2. In some studies, it is possible to obtain multiple observations from each study participant. In general, studies of this sort in which *N* represents the number of observations will have more statistical power than studies involving the same number of data points in which only one observation is obtained from each participant. This phenomenon is discussed in more detail in Chapter 8.